WWW 画像検索システムをにおける有害画像フィルタリング手法
A Method of Filtering Hazardous Images on WWW Image Search Systems

中川 嘉之 † **獅々堀 正幹** † **柘植 覚** † **北 研二** ‡
Yoshiyuki Nakagawa Masami Shishibori Satoru Tsuge Kenji Kita

1. はじめに

近年，インターネットの普及に伴い，急速に Web サイト数が増大しているが，Web サイトの中には未成年者にとって不適切な情報が多く存在している．この問題に対処するため，利用者がアクセスできる情報を制限するフィルタリングシステムが開発されている [1][2].

特に，WWW 画像検索システムは，教育現場において資料収集のために頻繁に用いられているにもかかわらず，一般的なキーに対する検索結果内にも多くの有害な画像が表示されてしまうため，フィルタリング処理の適用が望まれている．現在，既存の WWW 画像検索システムでは，有害画像の URL をデータベース化することでフィルタリングするものも存在する．しかし，有効な URL がデータベース化されていないため，高精度なフィルタリングは実現できていない．

そこで本稿では，URL をパス毎に出現頻度を用いて重みづけを行うことにより，有害性の高い URL を部分的に識別しフィルタリングする手法を提案する．フィルタリングに必要な URL データベースは，数十個のキーワード群を用意しておくだけで自動で構築できる．

2. WWW 有害画像フィルタリングシステム

既存のフィルタリングシステムは大きく分類すると，(1) アクセス制限する URL 一覧をデータベース化する URL チェック方式，(2) アクセス制限するキーワード一覧をデータベース化するキーワードチェック方式に分けられる．以下に (1)，(2) の方式を用いたフィルタリングシステムの問題点を示す．

(1) Google や Yahoo が持つ独自のフィルタリングシステムでは，URL チェック方式が用いられている．URL チェック方式は有害画像の URL をデータベース化し，完全にマッチする有害画像の URL の対してのみフィルタリングを行う．両検索システムでは，有害画像の URL データベースが不十分なため精度が悪く，多くの有害画像が閲覧できてしまう．

(2) goo や AltaVista はキーワードチェック方式を用いており，有害な情報が検索されるであろう検索キーに対して制限を行っている．しかし，検索キーによっては多くの安全な画像に対しても制限してしまう．また，制限されていない検索キーに対しては制限されないという問題がある．

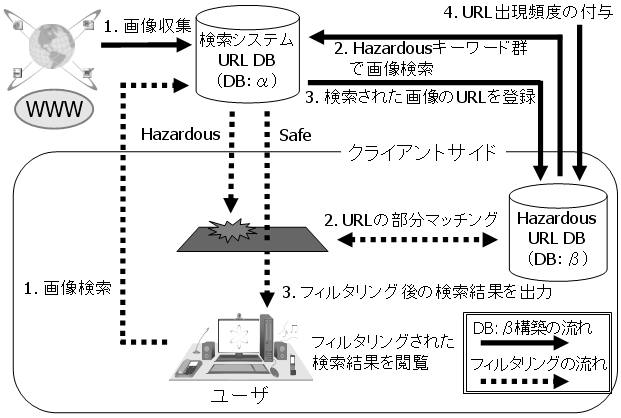

図 1: 本手法を用いたフィルタリングシステム

これらの問題から，高精度な有害画像のフィルタリングを行うためには，有害画像の URL をできるだけ多く網羅したデータベースを構築しなければならない．しかし，莫大な数の画像から有害画像を人手で判定するには多大な労力がかかる．また，URL データベースは定期的に更新されるため，有害画像の URL データベースは自動で構築できる方法が望ましい．そこで本稿では，数十個のキーワード群で自動構築できる URL データベースを用いて，未分類の URL を部分マッチングによりフィルタリングする手法を提案する．この手法を用いれば，有害画像の URL を完全に網羅しなくとも，データベースに登録された URL 数以上のフィルタリング効果を発揮することができる．

3. URL の部分マッチングによるフィルタリング手法

3.1 本手法を用いたフィルタリングシステム

本フィルタリング手法を用いることにより，図 1 に示すようなフィルタリングシステムを構築できる．このフィルタリングシステムは URL データベースの構築処理とフィルタリング処理に分けられる．以下に 2 つの処理についてそれぞれ説明する．

†徳島大学工学部
‡徳島大学高度情報化基盤センター

URL データベース構築処理：

既存の WWW 画像検索システムが持つ URL データベース（$DB:\alpha$）を用いて，有害画像の URL データベース（$DB:\beta$）を構築する．まず，WWW 画像検索システムに有害な画像を象徴するキーワード（Hazardous キーワード）群を入力して，検索される URL を $DB:\beta$ に登録する．その後，URL の部分マッチングに必要な情報として，$DB:\beta$ 中の各 URL 毎に URL の出現頻度を付与する．

フィルタリング処理：

ユーザが検索質問を入力すると，WWW 画像検索システムが出力する検索結果内の各 URL と $DB:\beta$ 中の URL との部分マッチングを行い，Hazardous と判定した URL にリンクを貼っている画像にアクセスできないようにする．このようにして，ユーザは有害画像がフィルタリングされた後の検索結果を閲覧することができる．

3.2 URL 出現頻度の抽出

$DB:\beta$ を用いて URL を部分的にマッチングさせるためには，$DB:\beta$ 中の URL の各部分毎に重みづけを行う必要がある．本手法では，URL の出現頻度を用いて重みづけを行った．以下に URL 出現頻度の抽出手順を示す．

手順 1： $DB:\beta$ 中の URL をパス毎に区切り，部分 URL（$PURL:\beta$）を抽出する．

手順 2： $DB:\beta$ 中の $PURL:\beta$ の出現頻度を求める．出現頻度が高いほど部分 URL の有害性が高いと言える．しかし，この情報だけでは単純に比較することはできないので，$DB:\alpha$ 中の $PURL:\beta$ の出現頻度を用いて正規化を行う．

手順 3： $DB:\alpha$ 中の $PURL:\beta$ の出現頻度を求める．この値は，検索システムの URL 検索機能を用いることで求めることができる．

手順 4： $DB:\alpha$ 中の $PURL:\beta$ の出現頻度を用いて，部分 URL 毎に正規化した $PURL:\beta$ の出現頻度（部分 URL の有害度 H_{url}）を求める．計算方法は以下の式 (1) で求める．

$$H_{url} = \frac{DB:\beta\text{中の } PURL:\beta\text{の出現頻度}}{DB:\alpha\text{中の } PURL:\beta\text{の出現頻度}} \quad (1)$$

これらの手順に対応した例を図 2 に示す．この例では 3 つの URL が $DB:\beta$ に登録されていると仮定し，それらの部分 URL 毎に H_{url} を求めている．

図 2: URL 出現頻度の抽出方法

3.3 URL の部分マッチング

部分 URL の有害度 H_{url} を用いて，検索結果中の有害画像に対しフィルタリングを行う．以下にフィルタリング方法の手順を示す．

手順 1： H_{url} の閾値 T を設定する．T を変動させることにより，フィルタリングのレベルを設定することが可能である．

手順 2： 検索結果中の画像にリンクする URL と，$DB:\beta$ 中の URL を URL の始端部から順に部分マッチングを行う．

手順 3： マッチした部分 URL 毎に H_{url} の判定を行う．T 以上となる部分 URL が $DB:\beta$ に登録されていれば，その部分 URL を含む URL を有害（Hazardous）とみなし，いなければその URL を無害（Safe）とみなす．

検索結果内の画像にリンクする全ての URL に対して部分マッチングを行い，Hazardous とみなされた URL

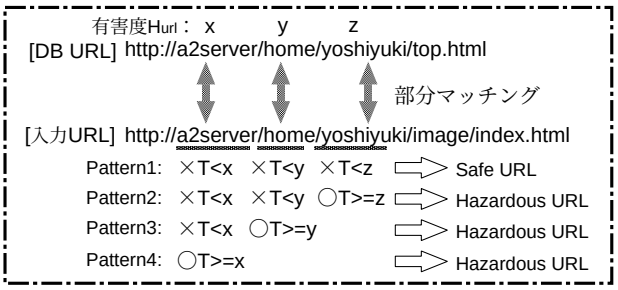

図 3: 部分マッチングによる有害な URL の判定

にリンクを貼っている画像を制限することで，検索結果内の有害画像をフィルタリングすることができる．部分マッチングによる本フィルタリング手法の例を図 3 に示す．本手法を用いれば，図のように URL の有害性を部分的に判定することができる．

4. Hazardous キーワードの選定

4.1 $DB : \beta$ 内に混入する Safe URL による影響

本フィルタリングシステムでは，Hazardous キーワードで検索された URL は，全て Hazardous な情報を含む URL と仮定して $DB : \beta$ に登録される．しかし，この URL の中には Safe なものも存在するため，部分マッチングで Safe な URL を誤って Hazardous と識別してしまう可能性がある．実際，Safe ありと Safe なしの $DB : \beta$ を作成して比較したところ，Hazardous 画像の正解率に平均 10%の差が見られた．また，キーワードによっては，Safe 混入の割合に大きな差があることが分かった．このことから，$DB : \beta$ 構築の際に，有害性の高い Hazardous キーワードのみを使用すれば，さらに高精度なフィルタリングができると考えられる．そこで，フィルタリングに有効な Hazardous キーワードを自動で選定する手法を考案した．

4.2 有効な Hazardous キーワードの選定手法

本選定手法は，Hazardous キーワード毎に検索結果内の画像にリンクする HTML 内のコンテンツを解析することにより，有害性の低いキーワードを除外する．図 4 に本選定手法の流れを示す．本手法では，Hazardous キーワード $key_i(1 \leq i \leq n)$ 毎に検索結果内の画像にリンクするページ $HTML_{ij}(1 \leq j \leq m)$ を取得し，$HTML_{ij}$ 内に出現する単語の内，検索キーワード以外の Hazardous キーワード $key_k(1 \leq k \leq n, i \neq k)$ の異なり数 $N(HTML_{ij})$ の平均値 $H(key_i)$ を以下の式 (2) で求める．この値を Hazardous キーワードの有害度と呼び，有害度の高い上位のキーワードのみを Hazardous

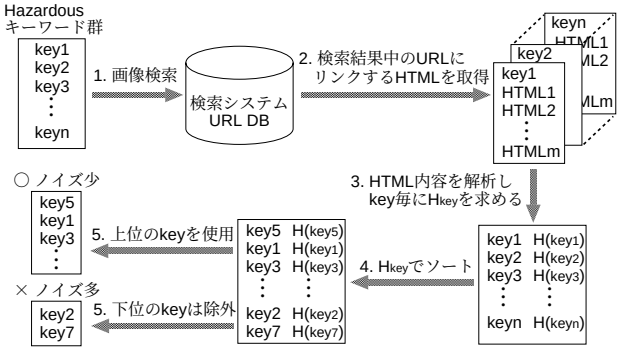

図 4: Hazardous キーワード選定手法

キーワードとして選定する．

$$H(key_i) = \frac{\sum_{j=1}^{m} N(HTML_{ij})}{m} \tag{2}$$

この手法を用いれば，Safe が多く混入している可能性が高い Hazardous キーワードを自動で除外することができる．

5. 評価実験

5.1 実験条件

Google Image Search を用いて 3 種類の $DB : \beta$ と評価用データを作成した．まず，54 個の Hazardous キーワードで検索し，検索結果上位 100 件の URL 計 4,189 件を $DB : \beta_{all}$ に登録した．次に，$DB : \beta_{all}$ 中から Hazardous なものだけを人手で取り出し，2,396 件の URL を $DB : \beta_{haz}$ に登録した．また，Hazardous キーワード選定手法を用いて，54 個の Hazardous キーワードの中から選定した上位 40 件のキーワードを用いて検索した 3,061 件の URL を $DB : \beta_{key}$ に登録した．最後に，Hazardous キーワードとは別に有害な画像が検索される可能性がある“ 看護婦 ”や“ 制服 ”といった 27 個の評価用キーワードで検索を行い，検索された URL 計 2,639 件を評価用データとした．更に，評価用データ中の URL を人手で Hazardous と Safe に分類し，456 件の Hazardous URL と 2,183 件の Safe URL を得た．

5.2 Hazardous 画像の再現率・適合率

評価用データに対して $DB : \beta_{all}$，$DB : \beta_{haz}$，$DB : \beta_{key}$ を用いてフィルタリングを行い，式 (3)，(4) に示す Hazardous 画像の再現率（R_{haz}），適合率（P_{haz}）を求めた．R_{haz} は評価用データ中の全 Hazardous 画像を正しくブロックできた割合を表し，P_{haz} はブロックした画像の中で本当に Hazardous 画像であった割合を表す．

$$R_{haz} = \frac{\text{正しく } Hazardous \text{ と分類された画像数}}{\text{全 } Hazardous \text{ 画像数}} \tag{3}$$

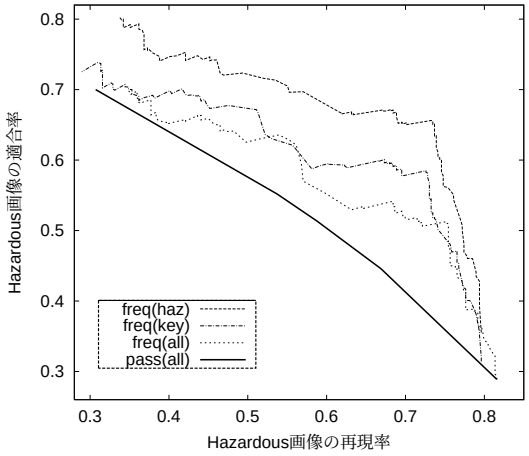

図 5: Hazardous 画像の再現率・適合率

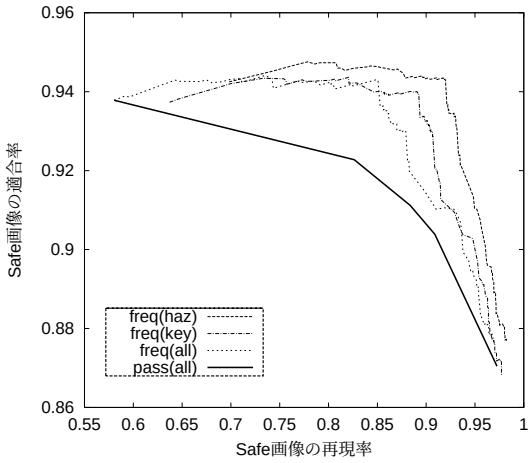

図 6: Safe 画像の再現率・適合率

$$P_{haz} = \frac{\text{正しく } Hazardous \text{ と分類された画像数}}{Hazardous \text{ と分類された画像数}} \quad (4)$$

閾値 T を 0.0〜1.0 まで 0.0001 毎に変化させた結果，図 5 に示す再現率・適合率曲線を得ることができた．図中の“ freq ”は本フィルタリング手法を用いた結果であり，“ pass ”は URL の出現頻度を考慮しない部分マッチングを用いた結果である．また，括弧内はフィルタリングに用いたデータベース名を示している．

5.3 Safe 画像の再現率・適合率

各 $DB : \beta$ を用いて評価用データに対してフィルタリングを行い，5.2 と同様にして Safe 画像の再現率（R_{saf}），適合率（P_{saf}）を求めた．R_{saf} は検索した Safe 画像に対してアクセスを許す割合を表し，P_{saf} はアクセスを許した Safe 画像の中で本当に Safe 画像であった割合を表している．Safe 画像の再現率・適合率曲線のグラフを図 6 に示す．

表 1: 各手法における F 尺度の平均値

	freq(haz)	freq(key)	freq(all)	pass(all)
haz	0.5992	0.5551	0.5498	0.5032
saf	0.8648	0.8469	0.8295	0.8243

5.4 F 尺度の平均値

各グラフにおいて，曲線毎に再現率・適合率の幅がまばらであり，曲線が交差する部分があるため，各手法の精度を順位付けするのは困難である．そこで，再現率と適合率を総合的な観点から 1 つの値により評価するために F 尺度を求めた．F 尺度は以下の式 (5) で求めることができる．

$$F = \frac{2RP}{R+P} \quad (5)$$

各グラフで，再現率を 0.0〜1.0 まで 0.05 毎に区切った計 101 点の F を計算し，その平均値を求めた．各手法毎の F の平均値を表 1 に示す．“ pass(all) ”に比べ“ freq(all) ”が高い値を示していることから，正規化した頻度を用いたフィルタリング手法が有効であるといえる．また，“ freq(all) ”に比べ“ freq(key) ”の値が向上しているので，キーワード選定を行った結果，$DB : \beta_{all}$ よりも Safe URL が少ない URL データベースの構築に成功しているといえる．最終的に，人手でレイティングした結果である“ freq(haz) ”に最も近い精度が“ freq(key) ”であり，本手法の有効性が確認できる．

6. まとめ

本稿では，URL をパス毎に重みづけを行うことにより，有害性の高い URL を部分的に識別しフィルタリングする手法を提案した．また，URL データベース構築に必要な Hazardous キーワードの選定手法を考案した．評価実験では，両手法を組み合わせると，人手でレイティングした結果に最も近い精度が得られ，本手法の有効性を確認できた．今後は，関連キーワード自動収集手法 [3] を用いた Hazardous キーワード拡張によるフィルタリング精度の向上を行いたい．

参考文献

[1] 井ノ上, 帆足, 橋本: 文書自動分類手法を用いた有害情報フィルタリングソフトの開発, 信学論 D-II, J84-D-II, No.6, pp.1158–1166, 2001.

[2] 武者, 広池, 森本, 松田: WWW 有害情報のフィルタリングのための画像判別手法, FIT 2002, I-82, pp163–164, 2002.

[3] 竹安, 獅々堀, 柘植, 北: 出現 URL の類似性に着目した WWW 空間からの関連キーワード自動収集手法, 言語処理学会第 11 回年次大会, 2005.